Discussion review 4

Math 181B

1 Review

1.1 χ^2 test for independence

- **Test statistics:** Suppose that n observations are taken on a sample space partitioned by the events $A_1, A_2, \ldots, A_r$ and also by the events $B_1, B_2, \ldots, B_c$. Let $p_i = P(A_i); q_j = P(B_j)$, and $p_{ij} = P(A_i \cap B_j), i = 1, 2, \ldots, r; j = 1, 2, \ldots, c$. Let X_{ij} denote the number of observations belonging to the intersection $A_i \cap B_j$. Then

  – the random variable

  $$D_2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - np_{ij})^2}{np_{ij}}$$

  has approximately a χ^2 distribution with $rc - 1$ degrees of freedom (provided $np_{ij} \geq 5$ for all i and j).

  – In practice, p_{ij} is usually **unknown**. To test H_0: the A_i 's are independent of the B_j 's, we need to estimate p_{ij} first. Then, we can calculate the test statistic

  $$d_2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(k_{ij} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j}$$

  where k_{ij} is the number of observations in the sample that belong to $A_i \cap B_j, i = 1, 2, \ldots, r; j = 1, 2, \ldots, c$ and $\hat{p}_i$ and $\hat{q}_j$ are the maximum likelihood estimates for p_i and q_j, respectively. The null hypothesis should be rejected at the α level of significance if

  $$d_2 \geq \chi^2_{1-\alpha,(r-1)(c-1)}$$

  (provide $n\hat{p}_i\hat{q}_j \geq 5$ for all i and j.)

- **Degree of freedom:** In general, the number of degrees of freedom associated with a goodness-of-fit statistic is given by the formula

  $$\text{df} = \text{ number of classes } - 1 - \text{ number of estimated parameters.}$$

  So for d_2, the number estimated parameters is $(r - 1) + (c - 1)$

- **Intuition:** If the events A_i and B_j are independent, then we have $P(A_i \cap B_j) = P(A_i)P(B_j)$ for all i and j. Under the null hypothesis, the expected number of observations that belong to $P(A_i \cap B_j)$ is np_ip_j; this number should be close to the true observation k_{ij}.

1.2 Linear regression

1.2.1 Simple linear regression

- **Formula:** Given n points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the straight line $y = a + bx$ minimizing

$$L = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

has slope

$$b = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2}$$

and an intercept term

$$a = \frac{\sum_{i=1}^n y_i - b\sum_{i=1}^n x_i}{n} = \bar{y} - b\bar{x}$$

- **Equivalent slope format:**

$$b = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{r s_y}{s_x};$$

where

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \;;\; s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \;;\; s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- **Remark about correlation r:**
  - Correlation r can only quantify the **linear** association.
  - Correlation is not equivalent to causation.
  - Uncorrelated two random variables may be dependent.

  (Advanced techniques, e.g., distance correlation, can quantify both linear and non-linear association.)

- **Some facts about simple linear regression**
  - The regression line always goes through $(\bar{x}, \bar{y})$.
  - If $x = \bar{x} + s_x$, i.e. x is one standard deviation above the mean, we predict y to be r standard deviations above the mean.

1.2.2 Nonlinear models

If our data present some non-linearity, we can do some transformation on data to create the linear relationship on transformed data:

- **Exponential regression:**

$$y = ae^{bx} \iff \ln y = \ln a + bx; \text{ for } y > 0$$

Then, we have

$$b = \frac{n\sum_{i=1}^{n} x_i \ln y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} \ln y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \; ; \; \ln a = \frac{\sum_{i=1}^{n} \ln y_i - b\sum_{i=1}^{n} x_i}{n}$$

- **Logarithmic regression:**

$$y = ax^b \iff \log y = \log a + b \log x; \text{ for } y > 0$$

We have:

$$b = \frac{n\sum_{i=1}^{n} \log x_i \cdot \log y_i - \left(\sum_{i=1}^{n} \log x_i\right)\left(\sum_{i=1}^{n} \log y_i\right)}{n\sum_{i=1}^{n} \left(\log x_i\right)^2 - \left(\sum_{i=1}^{n} \log x_i\right)^2} \; ; \; \log a = \frac{\sum_{i=1}^{n} \log y_i - b\sum_{i=1}^{n} \log x_i}{n}$$

- **Logistic regression:**

$$y = \frac{L}{1 + e^{a+bx}} \iff \ln\left(\frac{L-y}{y}\right) = a + bx; \text{ for } \frac{L-y}{y} > 0$$